PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation
AUTHORS	Pacurariu, Alexandra; Plueschke, Kelly; McGettigan, Patricia; Morales, Daniel; Slattery, Jim; Vogl, Dagmar; Goedecke, Thomas; Kurz, Xavier; Cave, Alison

VERSION 1 – REVIEW

REVIEWER	Marek Oleszczyk Jagiellonian University Medical College in Krakow, Poland
REVIEW RETURNED	01-Apr-2018

GENERAL COMMENTS	This is a very important paper to be published, improving information about availability and quality of Electronic healthcare databases. Few minor improvements might make it even more useful: 1. A few sentences on rationale behind inclusion criteria might be helpful. In lines 145-9 there are three of them stated, without any explanation - why those three (and no less, no more) were chosen. 2. The title of figure 1 is "Figure 1. Scoring of the usefulness of electronic healthcare databases available in Europe for the benefit-risk evaluation of medicines." and in this figure no scoring system is present. 3. Figure 3 - is difficult to read without more detailed plot and/or numbers; information on meaning of lines/boxes etc. (mean/median/max/min) might be repeated in the figure 4. The authors describe scoring system in the methods but there's no clear summary of the results is presented. Therefore no simple and clear information on the quality of reviewed databases is available. No conclusions on what score they consider as of good or bad quality for EHD.

REVIEWER	Peter Rijnbeek Department of Medical Informatics Erasmus MC Rotterdam, The Netherlands
REVIEW RETURNED	13-Apr-2018

GENERAL COMMENTS	General Comments This well written paper describes the usefulness of electronic healthcare databases (EHDs) for regulatory purposes. The authors have assessed publicly available information to identify eligible databases for this review. I have the following general comments: 1. Is the information that is captured from these databases enough to make a good assessment of their usefulness? Of course, it is important to know

	that they have access to drugs and events as a minimum requirement, but I wonder if only this information would be present for all databases in Europe this would answer EMA’s questions. I think we need to know much more about a database to know if it is suitable for a specific question. Should we not aim for sharing much more detailed information about the databases, i.e. standardized dashboard with information on drug use, age groups, number of events etc. It would be nice if the authors would discuss this in more detail. What is the ideal situation for EMA from regulatory perspective to have fast access for feasibility assessment and study execution? Moving to a CDM is a very important first step to achieve this. 2. The authors state correctly that a limitation of their study is that they may have missed data sources. I am certain this is the case, there are many more databases in Europe. This supports the need for a better registry of databases. The authors could consider to discuss this briefly in the paper. 3. The section and results on validation studies is limited. If available it would be interesting to discuss in more detail what these validation studies entailed, e.g. in which disease domains?, why?, how are these executed?, what was the gold standard etc. Related to that, if a database has performed a validation study for a certain outcome or exposure does this then really say something about the overall validity of the whole database? Furthermore, why is a database more valid if the ‘original’ patient records (what are those?) are reviewed by a medical professional? Minor comments Abstract Line 28. “EHDs are important but they are heterogeneous” I think the fact that there are heterogeneous databases is not by definition a bad thing. We need to have access to different types of data sources across Europe. I think this is referring to the structure, etc as specified later. I find it a bit strange that the Objective is the result of the study. Maybe better to say that the objective is the need to have more insight? Line 46. This sentence is not correct. Line 90. 53% and 31% does not read very well. New sentence? Line 94. Consider to move the 19 to line l93 Line 149. Suggest to remove the word ‘alone Line 174. Here is stated that information of laboratory tests results was not collected. Why is there then a statement made about those on line 308? Line 199. Data format is I think only referring to CDM or Not CDM. At this point in the text it is a bit vague what this is referring to.

	Line 269. Can the authors better specify what direct access means? Is this that EMA can buy/use the source data or are we also referring here to federated analyses which does not only has to be through third parties. This is also coming back at line 336. I think it is important to convey that a federated data analysis does not preclude regulatory questions. Line 363. The authors end the discussion with stating that Future work should focus on the existing validation studies and develop more robust validation measures. I personally think that Future work should focus on many other (maybe even more important) areas as I discuss in the general comments section above. Line 367. Is it far to say this is an "in-depth" evaluation? Would EMA not like to know much more than this (see general comment above) Table in supplementary data. IPCI does have Route of administration data.

VERSION 1 – AUTHOR RESPONSE

Manuscript (new title): Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation
Journal: BMJ Open
Initial decision: 16th April 2018

Reviewer 1

1. A few sentences on rationale behind inclusion criteria might be helpful. In lines 145-9 there are three of them stated, without any explanation - why those three (and no less, no more) were chosen.

> We have now expanded on the rationale for these selected criteria, which are mostly relevant for regulators. Please see the amended text in the methods section:
>
> 'These criteria were selected for their importance to regulators, and are key criteria for studies to meet regulatory requirements.'

2. The title of figure 1 is "Figure 1. Scoring of the usefulness of electronic healthcare databases available in Europe for the benefit-risk evaluation of medicines." and in this figure no scoring system is present.

> We actually did not perform a formal scoring but rather a coding of their characteristics. We do not think the overall score is relevant (therefore is not reported). Therefore we replaced the word 'scoring' with 'coding'.
>
> Figure 1. Coding of European electronic healthcare databases characteristics for the benefit-risk evaluation of medicines.
>
> We also inserted details about the coding system as a legend in Figure 1: 'The coding system was binary: 0 if information was absent and 1 if it was present. The degree of completion for a specific variable was not recorded. An exception to the binary classification was done for the accessibility

> variable: 0 - no access; 1 - indirect access through database owner or third party; 2 - direct access to specific data sources; 3 - direct access to full data source.’

3. Figure 3 - is difficult to read without more detailed plot and/or numbers; information on meaning of lines/boxes etc. (mean/median/max/min) might be repeated in the figure 4.

> We changed the legend of figure 3 to make it clearer.
>
> *Figure 3: European data sources and duration of data collection.*

4. The authors describe scoring system in the methods but there's no clear summary of the results is presented.Therefore no simple and clear information on the quality of reviewed databases is available. No conclusions on what score they consider as of good or bad quality for EHD.

> We did not perform a scoring but rather coding of their characteristics. Please also see answer to question 2.

Reviewer 2

1. Is the information that is captured from these databases enough to make a good assessment of their usefulness? Of course, it is important to know that they have access to drugs and events as a minimum requirement, but I wonder if only this information would be present for all databases in Europe this would answer EMA’s questions. I think we need to know much more about a database to know if it is suitable for a specific question. Should we not aim for sharing much more detailed information about the databases, i.e. standardized dashboard with information on drug use, age groups, number of events etc. It would be nice if the authors would discuss this in more detail. What is the ideal situation for EMA from regulatory perspective to have fast access for feasibility assessment and study execution? Moving to a CDM is a very important first step to achieve this.

> The reviewer is correct to state that ideally, much more information should be available in order to decide on usefulness of a data source for regulatory purposes. Such a detailed description is beyond the scope of this article, and can be better achieved in electronic repositories as EMIF, ENCePP database (free access) and Bridge to data (commercial access) which we now mention in the paper:
>
> ‘More detailed descriptions of database characteristics are provided in electronic repositories such as the European Medical Information Network (EMIF), the ENCePP resource database and the Bridge to Data initiative (21,22). However existing repositories are either incomplete, have a limited coverage or they require a fee for access, therefore restricting access to their information.
> This study helps identify databases with key characteristics as an entry door to further investigate with their owner their potential usefulness for a specific study.’

2. The authors state correctly that a limitation of their study is that they may have missed data sources. I am certain this is the case, there are many more databases in Europe. This supports the need for a better registry of databases. The authors could consider discussing this briefly in the paper.

> Although we might have missed some datasources, we think the main source of reduction in datasources in the fact that we restricted the analysis only to databases interesting from a regulatory perspective (e.g., accessible, complete, broad in scope and with longitudinal data capture). This lead to a very narrow sample, suggesting that the real availability from a regulatory perspective is lower. Examples of databases that were excluded:
>
> Prescription databases
> Registries
> Disease specific databases
> Local hospital databases
> Biobank resources

For context, the number of European databases listed in the (draft) EMIF database is 18 databases and in Bridge to Data is 117 (including registries and non-longitudinal datasets).

We also acknowledged this limitation in the discussion, please see below:

‘The difficulties we encountered when trying to map all the existing EHDs in Europe highlights again the need for more comprehensive and accessible repositories with EHDs.’

3. The section and results on validation studies is limited. If available it would be interesting to discuss in more detail what these validation studies entailed, e.g. in which disease domains?, why?, how are these executed?, what was the gold standard etc. Related to that, if a database has performed a validation study for a certain outcome or exposure does this then really say something about the overall validity of the whole database? Furthermore, why is a database more valid if the ‘original’ patient records (what are those?) are reviewed by a medical professional?

Indeed this section is limited, an independent validation of the datasources was considered out of scope and we relied only on the information provided by the database owners. We fully agree with the reviewer that the ‘overall validity of the database’ is a hard to prove status and that validation studies are usually outcome specific and study specific. We now added more details about the type of validation encountered for these databases and we expanded the discussion around the issue. We also did not consider the number of studies as an indicator of validity, but switched to a dichotomous variable: databases which have at least one study or none.

Please see below and the revised manuscript:

Results:

[….]

‘1.1. Validation studies

No published validation study was reported for 17 databases (50.0%), while a total of 42 validation studies were reported for the other 17 databases, with a median of 3 validation studies per database (range: 1-25). The validation concerned either specific health outcomes or prescription information. The most common gold standards used for the validation included paper based prescriptions, medical records, death records and perinatal deaths obtained from registries or national statistics reports. Some database owners have reported as validation studies the validation of prediction algorithms for various health outcomes as chronic kidney disease, ischaemic stroke and various types of cancers based on an estimating the absolute risk of a particular outcome in primary care patients with and without symptoms (1,2). It is debatable if these are truly validation studies according to our definition.’

Discussion

With regards to validation, 50% of databases had at least one validation study published. Validation should normally be done for the data elements collected for each study. The number of validation studies performed is not an indicator of the overall validity of the database but may inform researchers on the feasibility to perform study-specific validation in a specific database. A repository of validated outcomes in specific databases would reduce duplication of work. Such a repository should include a clear description of the methodology and limitations of the analysis.

Editor Comments to Author:

- Please edit the title so that it states the research question, study design and setting. This is the preferred format of the journal.

> We acknowledge the comment and changed the title to ‘Electronic healthcare databases in Europe: a descriptive study of their characteristics and usefulness for medicines regulators’

- Please complete and include a STROBE checklist, ensuring that all points are included and state the page numbers where each item can be found. The checklist can be downloaded from here: http://www.strobe-statement.org/?id=available-checklists

> We have attached the completed checklist. We would like to point out that our study is a descriptive study, therefore many of the items from the checklist are not applicable.

- Please include strength in your Strengths and Limitations section.

> We included the following strengths:
> - **Data extraction was based not only on publicly available information but complemented by information provided by database owners**
> - **The inventory was endorsed by a an expert working group, the ENCePP Working Group “Data Sources”**

VERSION 2 – REVIEW

REVIEWER	Marek Oleszczyk Department of Family Medicine Jagiellonian University Medical College, Poland
REVIEW RETURNED	10-Jun-2018

GENERAL COMMENTS	Thank you for answering previous concerns. There is also a major improvement of written English.

REVIEWER	Peter Rijnbeek Erasmus MC, Rotterdam, The Netherlands
REVIEW RETURNED	17-Jun-2018

GENERAL COMMENTS	I thank the reviewers for their replies to my questions and look forward to seeing this work published. One small remaining suggestion is to add if possible a better reference to the new sentence "the Spanish Information System for the Development of Research in Primary Care is implementing the model used in the ADVANCE project for vaccine studies (28)". It is now referring to the website which does not really help to understand the content of this model.

VERSION 2 – AUTHOR RESPONSE

Garcia-Gil M, Hermosilla E, Prieto-Alhambra D, Fina F, Rosell M, et al. (2011) Construction and validation of a scoring system for selection of high quality data in a Spanish population primary care database (SIDIAP). Inf Prim Care.19(3): 135–45.

ADVANCE Work Package 5 White Paper Proof-of-concept studies [Accessed 2018 May 23]. Available from : http://www.advancevaccines.eu/?page=publications&id=DELIVERABLES

The rest of the changes are just Editorial. I would also like to ask if we can move the statement regarding Patients Involvement, required by the journal and now at the end of Methods section, somewhere else at the end of the manuscript since does not appear to fit there. But I leave it to you to decide,